International Journal of
Environmental Research and Public Health

MDPI

Article

A Social Media Infodemic-Based Prediction Model for the Number of Severe and Critical COVID-19 Patients in the Lockdown Area

Qi Yan [1,2,*], Siqing Shan [1,2], Menghan Sun [1,2], Feng Zhao [1,2], Yangzi Yang [1,2] and Yinong Li [1,2]

[1] School of Economics and Management, Beihang University, Beijing 100191, China; shansiqing@buaa.edu.cn (S.S.); smh_gabriella@buaa.edu.cn (M.S.); zhao_feng@buaa.edu.cn (F.Z.); ymyang@buaa.edu.cn (Y.Y.); liyinong@buaa.edu.cn (Y.L.)
[2] Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operation, Beijing 100191, China
* Correspondence: bhyanqi@buaa.edu.cn

Abstract: Accurately predicting the number of severe and critical COVID-19 patients is critical for the treatment and control of the epidemic. Social media data have gained great popularity and widespread application in various research domains. The viral-related infodemic outbreaks have occurred alongside the COVID-19 outbreak. This paper aims to discover trustworthy sources of social media data to improve the prediction performance of severe and critical COVID-19 patients. The innovation of this paper lies in three aspects. First, it builds an improved prediction model based on machine learning. This model helps predict the number of severe and critical COVID-19 patients on a specific urban or regional scale. The effectiveness of the prediction model, shown as accuracy and satisfactory robustness, is verified by a case study of the lockdown in Hubei Province. Second, it finds the transition path of the impact of social media data for predicting the number of severe and critical COVID-19 patients. Third, this paper provides a promising and powerful model for COVID-19 prevention and control. The prediction model can help medical organizations to realize a prediction of COVID-19 severe and critical patients in multi-stage with lead time in specific areas. This model can guide the Centers for Disease Control and Prevention and other clinic institutions to expand the monitoring channels and research methods concerning COVID-19 by using web-based social media data. The model can also facilitate optimal scheduling of medical resources as well as prevention and control policy formulation.

Keywords: COVID-19; prediction model; machine learning; sentiment polarity; social media

Citation: Yan, Q.; Shan, S.; Sun, M.; Zhao, F.; Yang, Y.; Li, Y. A Social Media Infodemic-Based Prediction Model for the Number of Severe and Critical COVID-19 Patients in the Lockdown Area. *Int. J. Environ. Res. Public Health* **2022**, *19*, 8109. https://doi.org/10.3390/ijerph19138109

Academic Editors: Paul B. Tchounwou and Giuseppe La Torre

Received: 13 May 2022
Accepted: 28 June 2022
Published: 1 July 2022

1. Introduction

Considering the COVID-19 treatment practices in countries such as China, Italy, Spain, the United Kingdom, and the United States, severe and critical patients with COVID-19 in lockdown areas have relatively high mortality rates. Improvements in the recovery rate of the numbers of severe and critical COVID-19 patients is a critically important factor for effective control of the epidemic. The effective treatment of severe and critical COVID-19 patients requires massive medical resources, including setting up intensive care units (ICU) wards, equipping professional equipment and facilities (e.g., ventilators, EMCO), and deploying technical medical personnel. The increasing numbers of severe and critical COVID-19 patients are of great uncertainty, significantly disrupting the normal clinical treatment order of hospitals. If the allocation of medical resources is not timely and adequate, the regular and orderly treatment of patients will be seriously interrupted. Accurate prediction of the number of severe and critical COVID-19 patients helps to solve problems such as disordered treatment, unreasonable deployment of medical personnel, and inability to arrange professional medical equipment and facilities in advance. Therefore,

patient-oriented rescue measures can improve the effectiveness of treatment and help policymakers formulate tailored policy measures for epidemic prevention and control. Accurate prediction of the number of severe and critical COVID-19 patients is an urgent issue to combat the epidemic of COVID-19.

The purpose of this study is to predict the number of COVID-19 severe and critical patients in a lockdown area (Hubei Province, China) via data from the social media platform, Sina Weibo. Social media, also named social sensor, has the advantages of timeliness, openness, and easy access. Nguyen et al. (2020) studied the influence of customer emotions transited in social media on the investment strategies of investors and enterprise value [1]. He et al. (2018) investigated the way non-local enterprises utilized social media to promote effective purchases and to improve customer relationships after a crisis in public relations [2]. Severe and critical COVID-19 patients in the lockdown area are also popular topics on social media. People post their needs and concerns on social media, so that the "infodemic" quickly generates and spreads. Such social media data can reflect valuable information on severe and critical COVID-19 patients in a timely and effective manner, including the number of patients and the emotional intensity of patients. Recognizing, understanding, and controlling the "infodemic" helps to predict the development of the epidemic. This study uses social media data to build a prediction model with a forecast horizon, to predict the number of severe and critical COVID-19 patients in the lockdown area. The prediction model is capable of solving the predict uncertainty issues, and the forecast horizon of the model can save time for medical institutions, medical staff, and policymakers in advance to properly predict the disease epidemic and formulate prevention measurements.

This study contributes to the existing literature in three aspects. First, in terms of the research perspective, this study uses social media data to predict the number of severe and critical COVID-19 patients in the lockdown area, which is also a popular topic on social media. Second, in terms of research data, this study uses both real-time social media data and published consensus data on severe and critical patients for making predictions. Social media data are advantageous as they involve large numbers of people's sensors, provide timely information, and facilitate intense public discussion and public attention. Third, in terms of the research method, it proposed an improved machine learning prediction model based on the Hidden Markov model (HMM). This model can accurately predict the number of severe and critical COVID-19 patients in the lockdown area. Currently, more studies are focusing on exploring and predicting the severity of the COVID-19 epidemic. For example, Menni et al. (2020) used a symptom tracker to predict the number of potentially infected people by establishing a logistic regression model. COVID-19 symptom data in their study was collected through a smartphone app [3]. Yan et al. (2020) proposed a logically straightforward and easy decision rule to predict patients at the highest risk. Data were sourced from a blood sample database of COVID-19 infected patients in Wuhan, China [4]. Based on an exponential growth model, Tsang et al. (2020) estimated the impact of different definitions of COVID-19 confirmed cases on the epidemic's curve and transmission parameters and further predicted the number of new infections [5]. Based on a stochastic propagation dynamics model, Kucharski et al. (2020) predicted the dynamics of the outbreak and the future downward trend of the COVID-19 epidemic [6]. Yang et al. (2020) constructed an improved Susceptible-Exposed-Infectious-Removed (SEIR) model and Long-Short-Term-Memory (LSTM) neural network to predict the development trend of the COVID-19 epidemic under public health policy interventions [7]. The study used data from confirmed diagnoses published daily in China and travel data for public transportation. Based on consensus data of confirmed cases, Leung et al. (2020) proposed the Susceptible–Infectious–Recovered model to study the impact assessment of the dissemination and severity of COVID-19 under public health interventions outside Hubei, China [8]. Based on Wuhan's real-time mobility data and detailed consensus data of COVID-19 infection cases, Chinazzi et al. (2020) explored the role of imported COVID-19 infection cases in the spread of Chinese cities and measured the

impact of lockdown measures on the COVID-19 epidemic; their study results showed that the interventions implemented in China achieved plausible progress in slowing down the spread of COVID-19 locally [9]. Tian et al. (2020) also investigated the spread and control of the COVID-19 epidemic with data including case reports, human activities, and public health interventions; they found that the national emergency response appeared to combat the growth of the COVID-19 epidemic in China, to a large extent, and limit the scale of the epidemic, avoiding hundreds of thousands of potential cases by February 19 (day 50) [10]. In addition, Jia et al. (2020) measured the risk of regional outbreaks from the perspective of population mobility using mobile phone data [11]. By collecting clinical data from two New York hospitals, Cummings et al. (2020) studied epidemiological characteristics and evolution of severe and critical patients [12]. Álvarez-Mon et al. (2021) created an individualized analysis model of the risk of ICU admission or death for COVID-19 patients as a tool for the rapid clinical management of hospitalized patients in order to achieve resilience of medical resources [13]. To predict the time of the outbreak peak and the ICU beds required, Moghadas et al. (2020) simulated a COVID-19 outbreak by parameterizing U.S. demographics [14]. The results showed that the ICU capacity was insufficient in responding to this rapid outbreak. Policies that encourage self-isolation, may delay the peak of the epidemic and provide a time window for emergency mobilization to expand hospital capacity. Mohsin et al. (2021) identified several vital lifestyles and comorbidity-related risk factors of severe-critical COVID-19 [15]. HMM has gained growing popularity in the fields of bioinformatics, fault diagnosis, and disease prediction. Based on HMM, Perveen et al. (2019) used primary electronic health care cases to predict the risk of developing diabetes within 8 years and identify various factors affecting the risk of disease [16]. Barra et al. (2020) proposed a method for quantifying the number and duration of migraine attacks based on the patient's diary [17]. This model enables researchers to procure data with high inter-study validity. In addition, social media data have become one of the important data sources for scientific research [18], and the significant amount of hidden but valuable information it contains is worth mining and analyzing. Deng et al. (2018) comprehensively investigated the relationship between microblog sentiment and stock returns at the market and individual stock levels [19]. Roy et al. (2020) constructed an algorithm model called the "Suicide Artificial Intelligence Prediction Heuristic", to predict the future risk of suicidal thoughts by analyzing public Twitter data [20]. Gan et al. (2022) studied differences in posting patterns and user behaviors between men and women during the COVID-19 epidemic through social media [21].

2. Materials and Methods

2.1. Data Description

This paper aims to predict the number of perceived severe and critical COVID-19 patients from web-based social media data combined with officially published data and improve the predictions of severe and critical patient numbers in the lockdown area. The lockdown in Hubei Province began on 23 January 2020, and ended on 24 March 2020. Hubei Province began publishing consensus data on severe and critical patients on 20 January 2020, which are taken from public datasets that are freely available online. Therefore, the period for data collection in this study was from 0:00 on 20 January 2020 to 24:00 on 24 March 2020. To measure the prevalence of COVID-19 on social media, we collected and analyzed social media data through Sina Weibo (one of the largest web-based social media platforms in China) API (Application Programming Interface). We purchased the data collection service from Gooseeker. Gooseeker is an authorized API of Sina Weibo. Specifically, this study collected social media data that contains at least one of the following keywords including "FEVER", "GETTING HEAT", "INFECT", "PNEUMONIA", "COUGH", "VIRUS", and "ISOLATION". The post should be released from 20 January 2020 to 24 March 2020, by a blogger that is registered in Hubei Province. The whole data set amounts at 1,076,174 items, and each social media text contains information such as content, blogger name, posting time, and URL. The geographical distribution of social media data is shown in

Figure 1. Social media data contains COVID-19-related microblogs published by two types of patients: bloggers who already knew about self-infection and others who experienced similar COVID-19 symptoms but were unsure about infection. As a social sensor, social media data can reflect the characteristics and emotions of COVID-19 patients or potential patients in reality.

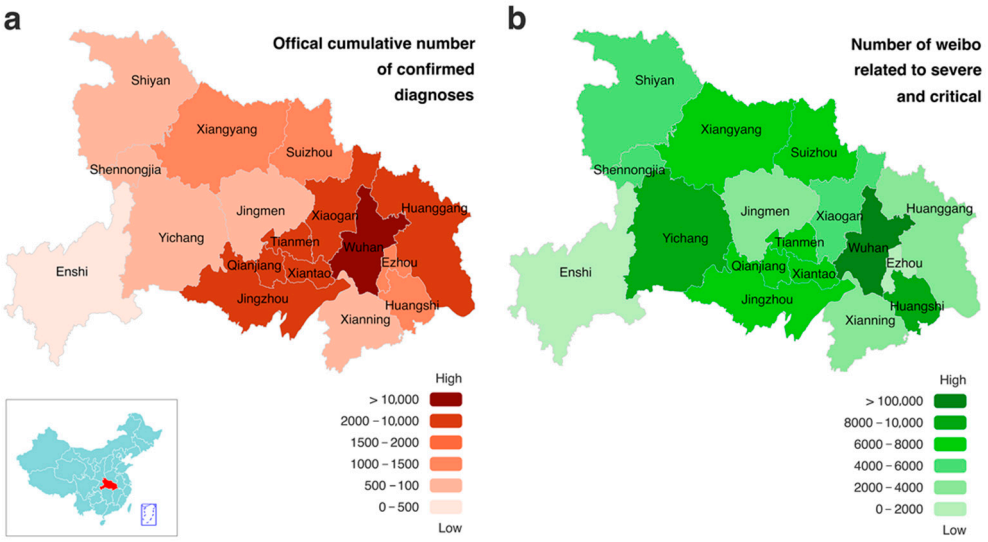

Figure 1. Geographic distribution of the number of perceived severe and critical COVID-19 patients and the number of confirmed cases by region in Hubei Province. Note: (**a**) Number of confirmed COVID-19 patients in Hubei Province during the lockdown period reported by consensus statistics. (**b**) Number of severe and critical patients related social media texts (abbreviated as SC-related social media texts) in Hubei Province during the lockdown period. The darker colors in the map indicate a larger number of COVID-19 patients in the region, indicating a more severe epidemic and a higher number of confirmed patients. In both maps, Wuhan is the darkest region, indicating that the epidemic situation shown by the number of SC-related social media texts is highly consistent with the actual situation. In Shiyan, Xiangyang, Suizhou, Jingmen, Enshi, Xianning, Huangshi, and other cities, there is a correlation between the number of SC-related social media texts and the cumulative number of confirmed COVID-19 patients, which shows a similar color change trend in the map; this indicates that the geographical distribution of the number of SC-related social media texts can effectively reflect the severity of COVID-19 in different regions. The number of SC-related social media texts in Xiaogan, Huanggang, Jingzhou, and other regions reflects the degree of the epidemic situation to be lower than the actual diagnosis, and higher than the actual diagnosis in Yichang. The trend difference in some cities may be due to inconsistencies in the number of COVID-19 patients and severe and critical COVID-19 patients. Since the information on geographic location collected from social media only contains 13 prefecture-level cities, data for the three county-level cities of Xiantao, Qianjiang, and Tianmen are considered as one area when merged with Jingzhou, and data for the Shennongjia forest area are considered as one area when merged with Shiyan.

2.2. Research Model

Based on HMM, this study builds a supervised machine learning model by considering the consensus data of severe and critical COVID-19 patients, the number of perceived severe and critical patients from social media, and the sentiment polarity of COVID-19-related social media texts; this model can realize the phased and horizontal prediction of severe and critical patient numbers in the lockdown area. The innovation of this study lies in three aspects. First, we construct an improved prediction model based on HMM. Many studies [22–25] have shown that machine learning works better in analyzing and processing nonlinear data. Compared with the time series model and regression model, HMM is a data-driven algorithm and has a substantial capacity to approximate nonlinear functions which can be used to satisfactorily predict nonlinear, time-varying data; moreover,

it is not necessary to choose the form of function artificially in HMM, which helps to explore the rules that are difficult to be found in the traditional linear regression model. Therefore, HMM can better adapt to the characteristics of complex data structure, which is consistent with our data feature in this paper; this machine learning model can realize the phased and prospective prediction of COVID-19 severe and critical patients in the lockdown area. The forecast horizon in the model guarantees more time for clinical treatment preparation. The model helps to make multiple predictions at various time points during the lockdown period. Second, by mining COVID-19-related information in social media data, we identified two key prediction indicators driven by social media data: the number of perceived severe and critical COVID-19 patients and the sentiment polarity of COVID-19-related social media texts. Compared with the studies that used official published data, it was verified that these two indicators improve the accuracy of the prediction of quantity. Third, we propose a COVID-19 prevention and control method, which can help the medical system to formulate treatment plans by predicting the number of severe and critical patients, dispatching medical personnel, and formulating prevention and control policies to enhance the quality of work; it can assist medical staff to perform early detection and intervention on patients, effectively reducing the case fatality rates as a result. Thus, this study provides a theoretical and methodical framework to utilize social media data in the prevention and control of the COVID-19 epidemic.

Note: Figure 2 depicts the four steps of constructing the two crucial prediction indicators of $NPSCP_t$. and $TSPI_t$. The first step is data preprocessing and the acquisition of seed words. The second step is to extract and expand the feature dictionary of COVID-19 severe and critical patients. The third step is to calculate the number of perceived severe and critical patients and the output $NPSCP_t$. The fourth step is to calculate sentiment polarity and the output $TSPI_t$ and $TSPI_t^{SC}$. The specific process of each step is shown in Figures 3–6.

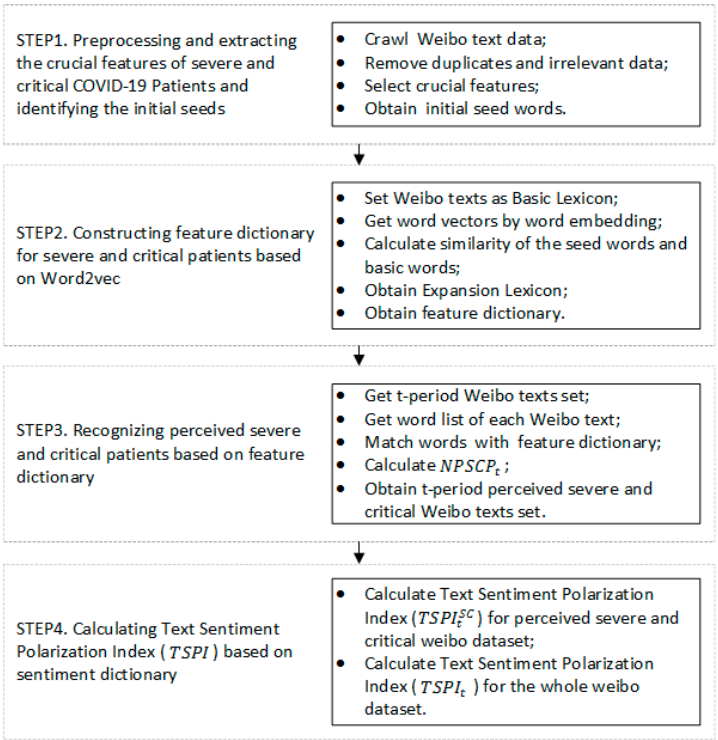

Figure 2. The conceptual framework of the integrated model for $NPSCP_t$, $TSPI_t$ and $TSPI_t^{SC}$.

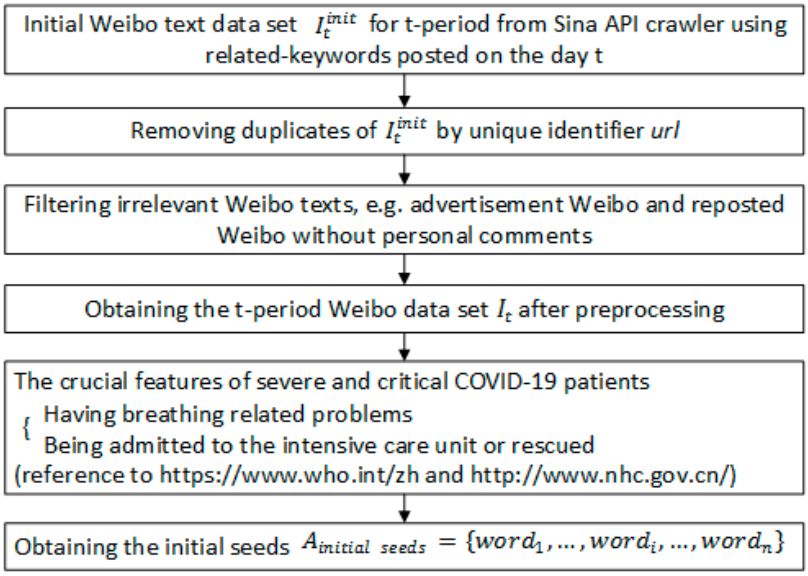

Figure 3. The first step of the integrated model.

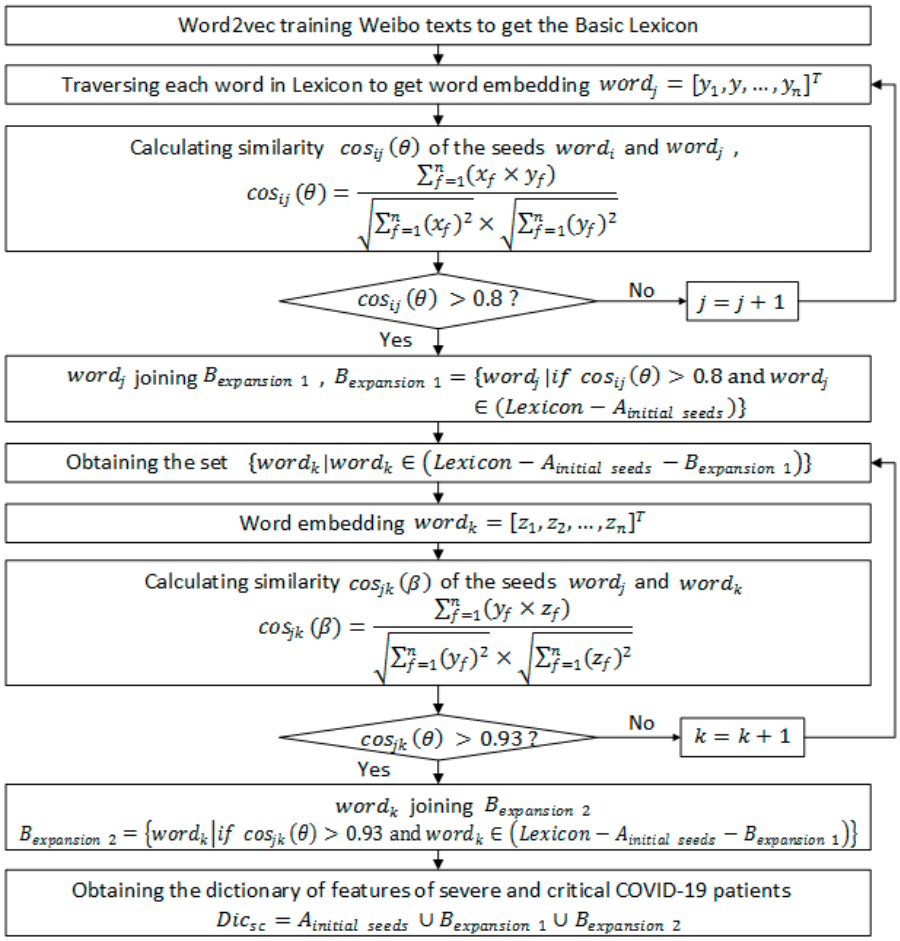

Figure 4. The second step of the integrated model.

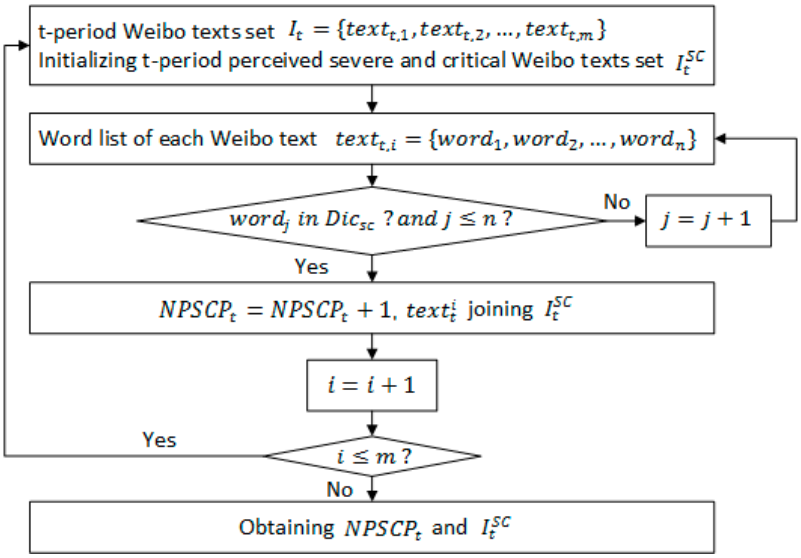

Figure 5. The third step of the integrated model.

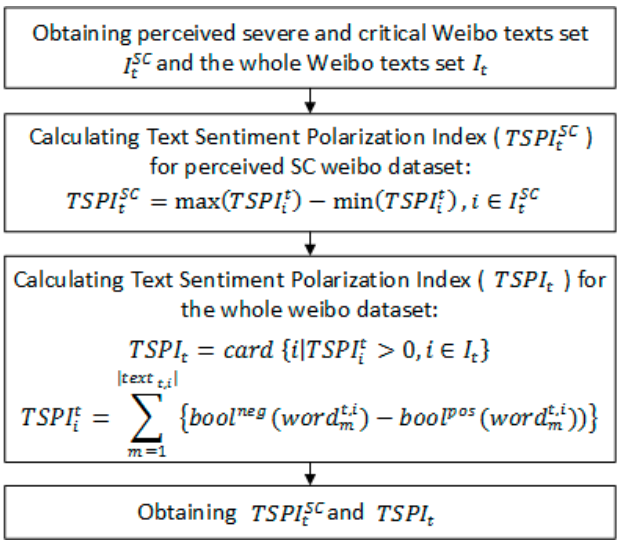

Figure 6. The fourth step of the integrated model.

NPSCP is the abbreviation for the number of perceived severe and critical patients in social media data. $NPSCP_t$ is defined as *NPSCP* of the *t*-period. The severe and critical patients are identified based on the constructed feature dictionary. When calculating the number of severe and critical social media texts in the *t*-period, the specific logic is that if there is a word in the social media text that matches any of the words in the feature dictionary, the social media text is recognized as a severe and critical text published by the patient or related parties. The detailed calculation process is as follows:

$$NPSCP_t = \sum_{i=1}^{m} (bool^{sc}(text_{t,i})) \quad (1)$$

$$bool^{sc}(text_{t,i}) = \begin{cases} 1, & \exists word_j \ in \ Dic_{sc} \\ 0, & \nexists \ word_j \ in \ Dic_{sc} \end{cases} \quad (2)$$

where I_t is the set of t-period social media texts; $text_{t,i}$ is the set of all words contained in each social media text; $bool^{sc}(text_{t,i})$ is a Boolean function, which takes values between $\{0, 1\}$. If any word in $text_{t,i}$ is included in Dic_{sc}, the value takes 1. If there is no word in $text_{t,i}$ included in Dic_{sc}, the value takes 0. The detailed calculation process is shown in Figure 2; this model proposes to use social media data to identify potential severe

and critical COVID-19 patients; it extracts the feature of the patients through social media sensors and screens out the implied COVID-19 severe and critical patients by user-generated content.

$TSPI$ refers to Text Sentiment Polarity Index, specifically $TSPI_t^{SC}$ and $TSPI_t$ respectively, representing text sentiment polarity index of social media texts perceived severe and critical related and text sentiment polarity index of all social media texts in the t-period. After data cleaning and recognition of social media texts related to perceived severe and critical patients, we calculated $TSPI_t^{SC}$ and $TSPI_t$ based on a sentiment dictionary National Taiwan University Semantic Dictionary (NTUSD). NTUSD is a sentiment dictionary based on Simplified Chinese, and it is widely used in sentiment analysis and sentiment modeling in Chinese scenes; it contains 2810 positive words and 8276 negative words. The calculation of $TSPI_t^{SC}$ and $TSPI_t$ is as follows:

$$TSPI_t^{SC} = max\left(TSPI_i^t\right) - min\left(TSPI_i^t\right), i \in I_t^{SC} \tag{3}$$

$$TSPI_t = card\left\{i \middle| TSPI_i^t > 0,\ i \in I_t\right\} \tag{4}$$

$$TSPI_i^t = \sum_{m=1}^{|text_{t,i}|}\left\{bool^{neg}\left(word_m^{t,i}\right) - bool^{pos}\left(word_m^{t,i}\right)\right)\right\} \tag{5}$$

where, $text_{t,i} = \left\{word_1^{t,i}, word_2^{t,i}, word_3^{t,i}, \ldots, word_{|text_{t,i}|}^{t,i}\right\}$ represents a certain word list of social media text in t-period after segmentation, I_t is for the whole social media data set in the t-period, I_t^{SC} is for the severe and critical related social media data set in t-period and function card calculates the cardinal number or size of a set. $bool^{neg}(word)$ and $bool^{pos}(word)$, represent the indication function of whether word w is in the NTUSD dictionary. If word w in $NTUSD_{neg}$, then $bool^{neg}(word)$ equals 1, else 0; it is the same for the value of $bool^{pos}(word)$. The detailed calculation process is shown in Figure 2; this model proposes the usage of social media text data to measure and profile the negative polarity of Weibo users, especially potential severe and critical COVID-19 patients, to help describe the severity of illness and future trends from a psychological perspective.

$NSCP_{1,t}$ refers to the actual number of severe and critical patients from period 1 to t, $NPSCP_{1,t}$ is the number of perceived severe and critical patients through social media from period 1 to t, and $TSPI_{1,t}$ is the number of negative sentiment texts from the period 1 to t. Let state sequence $L = \{l_1, l_2, \ldots, l_T\}$ indicate the basic state of the number of critical patients in the lockdown area, observation sequence $O = \{o_1, o_2, \ldots, o_T\}$ indicate the number of severe and critical patients that can be observed, and T denote the length of the sequence [26]. The transition between states can be described as a transition matrix, so every element is defined as follows:

$$a_{rv} = P(l_{t+1} = q_v | l_t = q_r, NSCP_{1,t}, NPSCP_{1,t}, TSPI_{1,t}) \tag{6}$$

where a_{rv} represents the probability transition to state q_v at time $t+1$ under the condition of the state q_r at time t, $r = 1, 2, \ldots, N; v = 1, 2, \ldots, N$. Each element in the observation probability matrix can be expressed as $b_v(o_g) = P(o_t | l_t = q_v)$, which represents the probability of generating observation o_t under the condition of the state q_v at time t. We assume $b_v(o_g)$ follows a Gaussian mixture distribution. Let $\pi_r = P(l_1 | q_r)$ is the probability at the state q_r when $t = 1$. To capture different amplitude changes, we model the predicted observation sequence after period $t + h$ as:

$$NSCP_{t+h}^* = \underset{o_g}{argmax}\left[\sum_{r=1}^{N}\sum_{v=1}^{N}\alpha_g(r)a_{rv}b_v\left(o_g\right)\varphi_g(v)\right],\ \forall g = 1, 2, \ldots, h \tag{7}$$

where $\alpha_g(r) = P(o_1, o_2, \ldots, o_t, r_t = q_r | a_{rv}, b_v\left(o_g\right), P(r_1 = q_r))$ is the probability of forward calculation at time t of q_r, $\varphi_g(v) = P(o_{t+1}, o_{t+2}, \ldots, o_T | r_t = q_r, a_{rv}, b_v\left(o_g\right), P(r_1 = q_r))$ is the probability of backward calculation from time $t+1$ to T of observation sequence. In this work, $NSCP$, $NPSCP$, and $TSPI$ are included in the transition matrix, which helps further

understand the influence of multidimensional observation factors. The range changes of severe and critical patients' numbers in states can be randomly transferred and captured by the Markov process; this study introduces the Gaussian mixture distribution to calculate an observation matrix that can fully learn the influence of multidimensional observation. For example, the number of severe and critical patients and emotions on Weibo helps to capture the fluctuations in patient numbers.

2.3. Ethics Statement

This study was based on an analysis of official data released daily by Hubei Province, China, which are public datasets that are freely available online and social media data from Sina API (Application Programming Interface). We purchased the data collection service from Gooseeker. Gooseeker is an authorized API of Sina Weibo. The data do not involve private information, such as personal name, gender, age, etc., and do not involve privacy and other related issues. All data can be used legally.

3. Results

Through a significant number of experiments, it showed that the improved prediction model proposed in this study has an outstanding performance after combining the two social media prediction indicators and the officially published data. The prediction performances and results comparison are shown in Figures 7–9. We conducted research to calculate the text sentiment polarity index, and present two indexes from two aspects, $TSPI_t^{SC}$ and $TSPI_t$. The two indexes are subject to distinct distributions and have different performances on prediction. The definition of $TSPI_t^{SC}$ is based on negative polarity, which indicates the sentiment polarity of each social media text; thus, its trend and distribution fluctuate with $NSCP_t$. In the peak period of the epidemic, its fluctuation ranges are large, whilst in other periods, its fluctuation ranges are relatively small. The daily distribution of severe and critical patients related texts' negative polarity is shown in Figure 10. When using $TSPI_t$ as a measure of sentiment polarity, prediction performance is greater than with $TSPI_t^{SC}$. Specifically, when using $TSPI_t$ as a polarity measure, the RMSE is 198.48, compared with the RMSE of 301.00 when using $TSPI_t^{SC}$ with the same parameters. As shown in Figure 11, the trend of $TSPI_t$ is not consistent with the trend of $NSCI_t$, which is evident from the low Pearson correlation coefficient (-0.31957) between them. Compared with $TSPI_t$, the Pearson correlation coefficient between $TSPI_t^{SC}$ and $NSCP_t$ is higher (0.61314), but $TSPI_t^{SC}$ has poorer performance in prediction.

Note: Table 1 shows the prediction results in the forecast horizon. After adding the two indicators of perceived severe and critical patients and the text sentiment polarity index, the RMSEs are gradually decreasing and the predicted values are getting closer to the actual values. Furthermore, this study adjusted the number of training days to conduct comparative experiments and verify the robustness of the model.

Table 1. RMSEs with different training days and HMM observed variables.

Number of Days	HMM Observed Variables	RMSE
65	$NSCP$	422.41
	$NSCP + NPSCP$	251.55
	$NSCP + NPSC + TSPI$	198.48
60	$NSCP$	633.16
	$NSCP + NPSCP$	462.65
	$NSCP + NPSC + TSPI$	369.96
55	$NSCP$	838.38
	$NSCP + NPSCP$	576.18
	$NSCP + NPSC + TSPI$	500.25

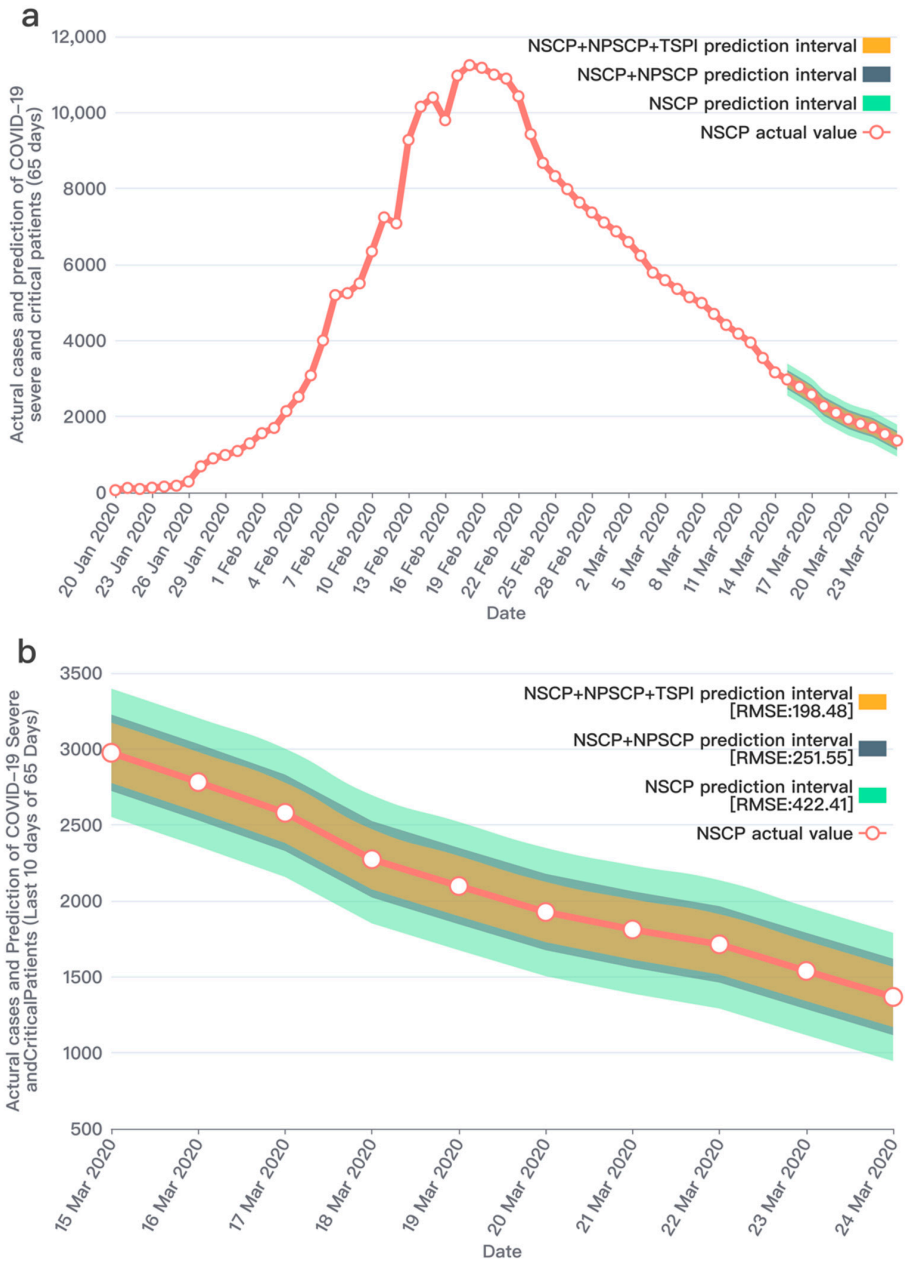

Figure 7. Prediction of the number of severe and critical COVID-19 patients (65 days data set). Note: (**a**) The entire prediction of 65 days during the lockdown period. The first 55 days are the training set and the last 10 days are the test set. The red line with circled points represents the actual values of the number of severe and critical patients; The green area represents the prediction interval obtained from the training data only including $NSCP_{1,55}$, which is based on the red line, RMSE as the fluctuation range. The blue area represents the prediction interval obtained from the training data including $NSCP_{1,55}$ and $NPSCP_{1,55}$, which is based on the red line, RMSE as the fluctuation range. The yellow area represents the prediction interval obtained from the training data including $NSCP_{1,55}$ and $NPSCP_{1,55}$, and $TSPI_{1,55}$, which is based on the red line, RMSE as the fluctuation range. (**b**) The 10 days prediction interval in 65 days. It can be found that the results predicted by the two sets of training data are better than the results of only one set because the RMSE with three sets is 198.48 which is evidently lower than the RMSE 422.41 with one set; this result shows that $NPSCP$ significantly contributes to the prediction. Furthermore, the prediction results (RMSE: 198.48) obtained from the three sets are better than those of the two sets (RMSE: 251.55), which shows that the $TSPI$ has a positive effect in predicting.

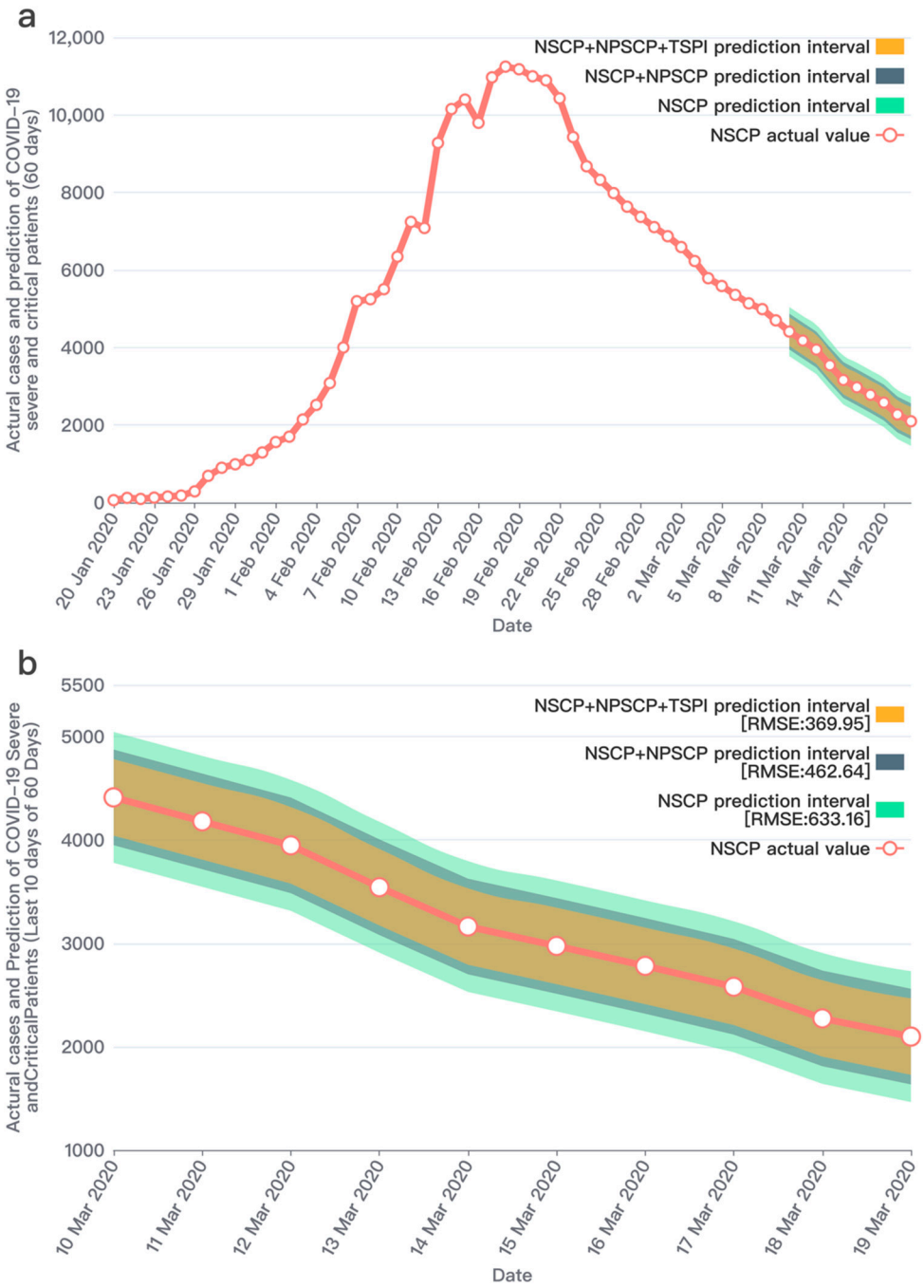

Figure 8. Prediction results of the number of severe and critical COVID-19 patients (60 days data set). Note: (**a**) The entire prediction of 60 days during the lockdown period. The first 50 days are the training set, and the last 10 days are the test set. The legend description is the same as Figure 7a. (**b**) The 10 days prediction interval in 60 days. From the prediction results, we can verify our conclusion that $NPSCI$ and $TSPI$ can help predict the number of severe and critical COVID-19 patients.

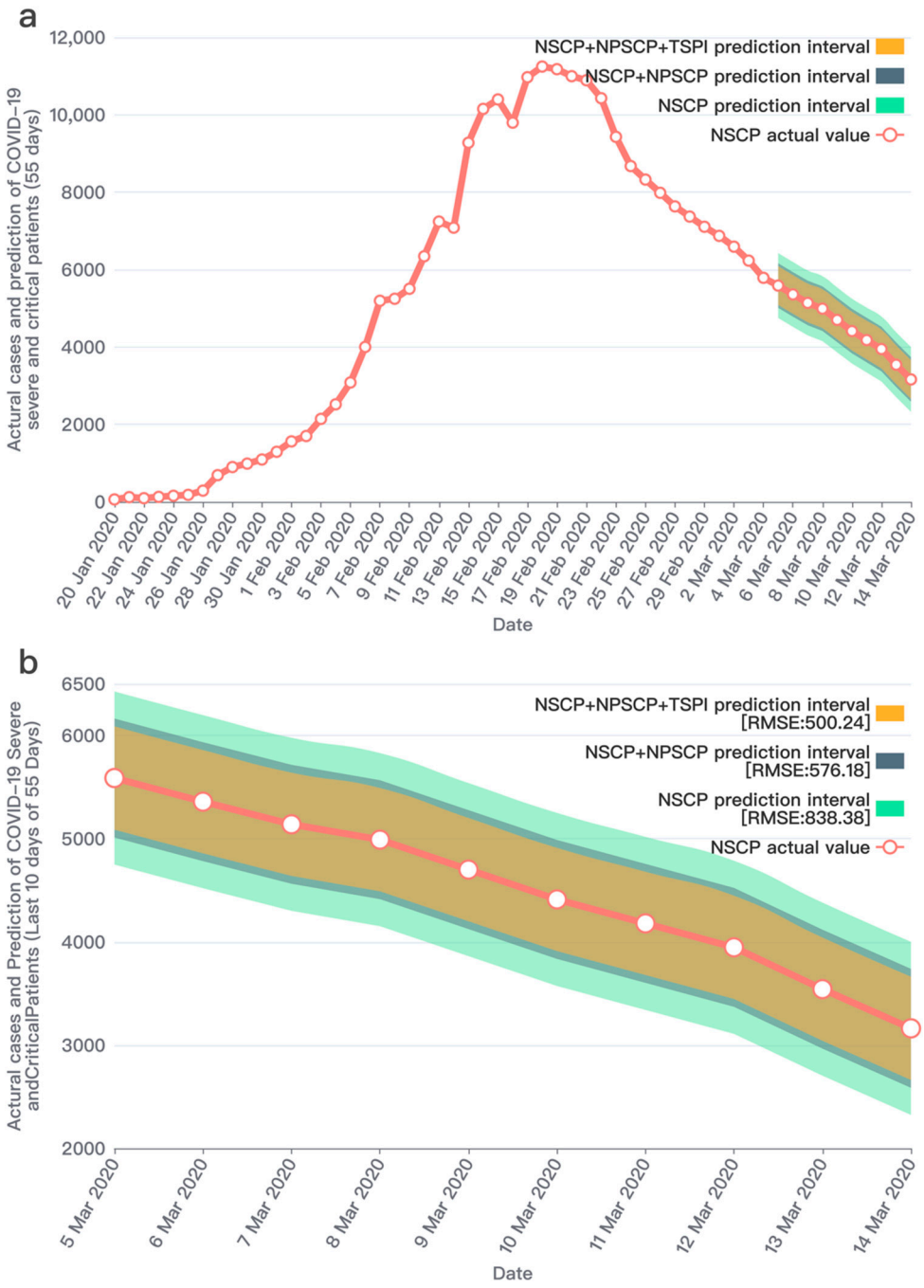

Figure 9. Prediction results of the number of severe and critical COVID-19 patients (55 days data set). Note: (**a**) The entire prediction of 55 days during the lockdown period. The first 45 days are the training set, and the last 10 days are the test set. The legend description is the same as Figure 7a. (**b**) The 10 days prediction interval in 55 days. From the prediction results, we can see that after adding *NPSCI* and *TSPI*, our prediction results still improved. Although it is not as good as the 65-day and 60-day performance, we analyze that it is related to limited samples in the training data set and insufficient machine learning.

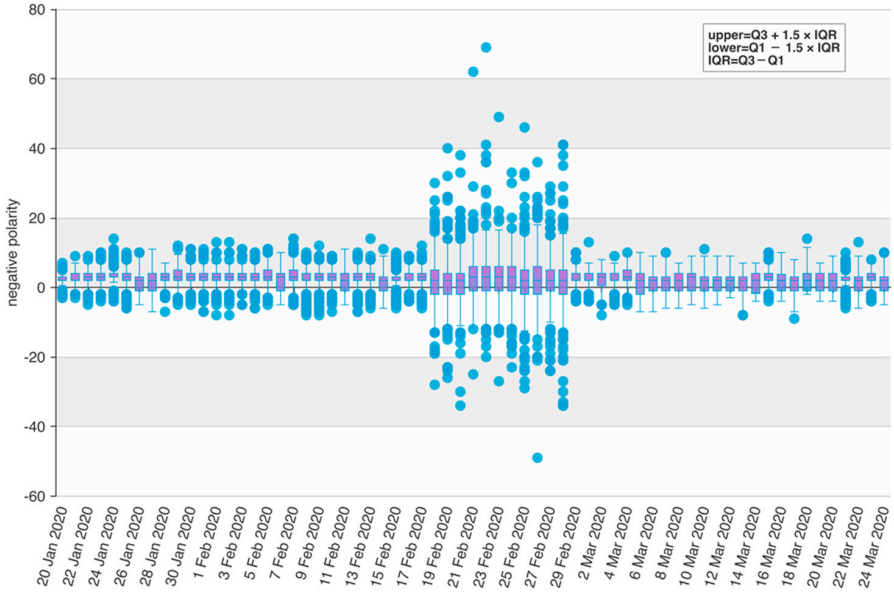

Figure 10. Negative polarity distribution of severe and critical related social media texts in the lockdown period (with outliers). Note: Negative polarity refers to the difference between the number of negative words and positive words in certain social media texts according to the sentiment dictionary NTUSD. During the period from 20 January to 17 February, the overall negative polarity of perceived severe and critical social media texts is negative, and the extreme values are at a relatively stable level with little fluctuation. There is a significant surge from 18 February to 28 February in SC-related negative polarity (the most obvious performance on 22 February), with the span and diversity of emotions increasing significantly, as the upper and lower values of emotions showed. After 28 February, with the gradual control of the epidemic, the overall negative emotions of SC-related microblogs show a gradual reduction trend.

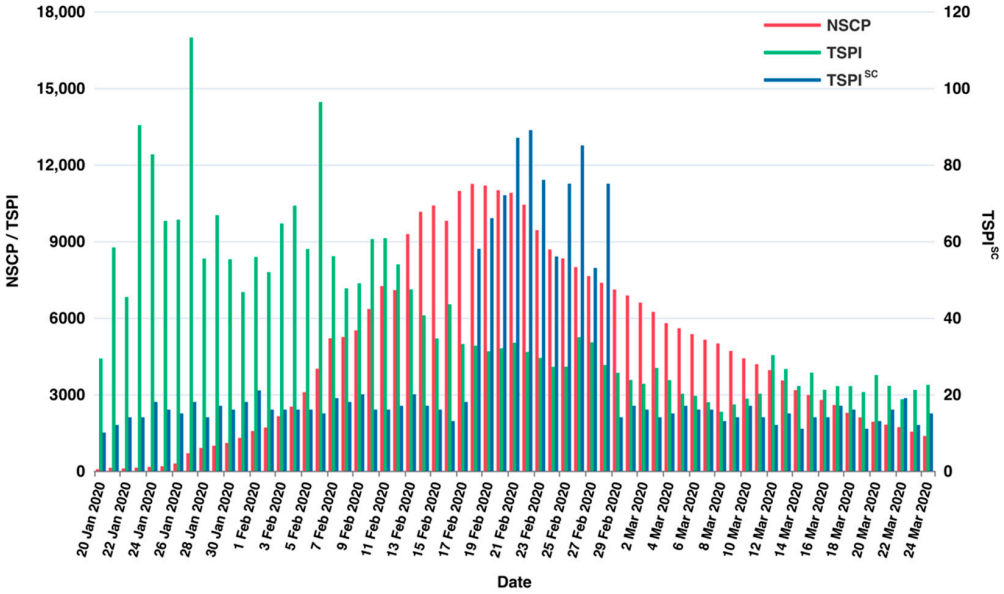

Figure 11. Trends of $NSCP$, $TSPI$ and $TSPI^{SC}$ during the lockdown period. Note: The trend of $TSPI$ started to rise before the city was locked down (23 January 2020). Both the trend and peak of $TSPI$ are earlier than those of $NSCP$, indicating that Weibo is an effective social sensor that can reflect the characteristics of COVID-19 patients or potential patients in reality. The peak lag bias of $TSPI^{SC}$ may be due to the following reason. Only after the government reported the exact number of COVID-19 patients to the public, did the news media begin to release news about the exact number of COVID-19 patients and the social media discussion also lags behind the release of the news.

This study reveals an interesting phenomenon. When the curves of actual data and perceived social media data are highly consistent, prediction performance is often inadequate. The consistency between the two data series can be described by a Pearson correlation coefficient, and the prediction performance can be manifested by the RMSE indicator. Specifically, a higher Pearson correlation coefficient indicates higher consistency and richness in current information, while a higher RMSE indicates poorer performance in prediction and lack of trend information; therefore, the phenomenon might imply that with abundant current information in perceived social media data and lacking trend information, prediction accuracy may be always adequate. On the contrary, when current information in perceived social media data is lacking but trend information is abundant, it may help to profile future trends and improve prediction performance; thus, this phenomenon suggests that the definition $TSPI_t$ contains information on future trends of $NSCP_t$, which facilitates the early perception of $NSCP_t$.

4. Conclusions

By using the social media infodemic to compute the number of perceived severe and critical COVID-19 patients and sentiment polarity, this study enables a more accurate prediction of patients for the horizon period in the lockdown area in comparison to a traditional consensus-based model. Social media data related to severe and critical COVID-19 patients are real-time and publicly available data sources that contain richer information than traditional hospital-reported consensus. Analysis of social media data reveals the reasons why some severe and critical patients did not seek medical treatment in the hospital, including preparing to visit the hospital but not taking action, expecting medical treatment but having no access due to limited hospital beds, having no intention for treatment, or not realizing that they needed treatment. Therefore, the number of perceived severe and critical COVID-19 patients explored through social media contains valuable information that cannot be obtained from traditional hospital consensus statistics. The sentiment polarity of perceived COVID-19 patients from social media includes the emotions concerning themselves, or those around them, or related information in the lockdown area. From the perspective of social sensor theory, social media data depicts the public emotions and opinions concerning COVID-19 patients. The potential evolution and trends in the number of COVID-19 patients were also revealed; this information contains possible trends in severe and critical COVID-19 patients. With a specified horizon perspective, the prediction model proposed in the study can predict the number of perceived severe and critical COVID-19 patients, the text sentiment polarity index, and the number of severe and critical patients in hospitals; this model has several advantages. First, it comprehensively reflects the actual number of severe and critical COVID-19 patients as it includes both the official published data and perceived severe and critical patients by social media. Second, the trend of severe and critical COVID-19 patients receive support from social media public opinion information (COVID-19 sentiment polarity). Further, COVID-19-related social media data can be obtained in real-time at a low cost. To a certain extent, it can complement the shortage of hospital consensus data. Importantly, this model has a specified horizon; this prospect period can increase the preparation time for hospital clinical treatment, medical personnel deployment, and epidemic prevention policy formulation. Theoretically, the prediction model proposed in this paper, which combines actual patient consensus data, social media-perceived patient data, and social media-perceived public sentiment data, is versatile and powerful for application in various research domains including public health and commercial sales; this approach provides a research paradigm based on social media infodemic for building real-time predictions and early warning models in the future.

Practically, the results of this prediction model can play a key role in COVID-19 prevention and control. It can: support hospitals to formulate a clinical treatment plan and optimize the operation of clinical resources; provide support for medical facility manufacturers and reserve companies in arranging production, storage, and transportation; support the Centers for Disease Control and Prevention and other government agencies to

formulate epidemic prevention and control policies and strengthen or liberate lockdown measures in epidemic areas; and provide practical guidance for building an epidemic prediction and early warning system based on network sensors.

When people post their emotions and opinions on COVID-19 on the Internet, they also provide rich information to understand the evolution of COVID-19. Data analysis and prediction technology constructed in this paper help to detect the developing trend of prior infectious diseases, and provide data support for further prevention and control. In terms of directions for future work, it is worth trying to develop further experiments to predict outbreaks in multiple lockdown areas and compare the epidemic situation with different characteristics.

Author Contributions: Conceptualization, Q.Y. and S.S.; Data curation, M.S. and Y.L.; Formal analysis, Q.Y. and M.S.; Funding acquisition, S.S.; Investigation, Q.Y. and Y.Y.; Methodology, Q.Y., S.S. and M.S.; Project administration, S.S.; Re-sources, Q.Y.; Software, S.S. and F.Z.; Supervision, S.S.; Validation, F.Z. and Y.L.; Visualization, Q.Y. and M.S.; Writing—original draft, Q.Y. and M.S.; Writing—review and editing, Q.Y., S.S. and F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 72071010 and by National Natural Science Foundation of China, grant number 71771010.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors gratefully acknowledge the support of Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations. Besides, the authors also thank the anonymous reviewers for insightful comments that helped us improve the quality of the paper.

Conflicts of Interest: The authors declare no conflict of interest. The authors declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

References

1. Nguyen, H.; Calantone, R.; Krishnan, R. Influence of Social Media Emotional Word of Mouth on Institutional Investors' Decisions and Firm Value. *Manag. Sci.* **2020**, *66*, 887–910. [CrossRef]
2. He, S.; Rui, H.; Whinston, A.B. Social Media Strategies in Product-Harm Crises. *Inf. Syst. Res.* **2018**, *29*, 362–380. [CrossRef]
3. Menni, C.; Valdes, A.M.; Freidin, M.B.; Sudre, C.H.; Nguyen, L.H.; Drew, D.A.; Ganesh, S.; Varsavsky, T.; Cardoso, M.J.; El-Sayed Moustafa, J.S.; et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat. Med.* **2020**, *26*, 1037–1040. [CrossRef]
4. Yan, L.; Zhang, H.-T.; Goncalves, J.; Xiao, Y.; Wang, M.; Guo, Y.; Sun, C.; Tang, X.; Jing, L.; Zhang, M.; et al. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2020**, *2*, 283–288. [CrossRef]
5. Tsang, T.K.; Wu, P.; Lin, Y.; Lau, E.; Leung, G.; Cowling, B.J. Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: A modelling study. *Lancet Public Health* **2020**, *5*, e289–e296. [CrossRef]
6. Kucharski, A.J.; Russell, T.W.; Diamond, C.; Liu, Y.; Edmunds, J.; Funk, S.; Eggo, R.M.; Sun, F.; Jit, M.; Munday, J.D.; et al. Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *Lancet Infect. Dis.* **2020**, *20*, 553–558. [CrossRef]
7. Yang, Z.F.; Zeng, Z.Q.; Wang, K.; Wong, S.S.; Liang, W.; Zanin, M.; Liu, P.; Cao, X.; Gao, Z.; Mai, Z.; et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *J. Thorac. Dis.* **2020**, *12*, 165. [CrossRef]
8. Leung, K.; Wu, J.T.; Liu, D.; Leung, G.M. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: A modelling impact assessment. *Lancet* **2020**, *395*, 1382–1393. [CrossRef]
9. Chinazzi, M.; Davis, J.T.; Ajelli, M.; Gioannini, C.; Litvinova, M.; Merler, S.; Piontti, Y.; Pastore, A.; Mu, K.; Rossi, L.; et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **2020**, *368*, 395–400. [CrossRef]
10. Tian, H.; Liu, Y.; Li, Y.; Wu, C.-H.; Chen, B.; Kraemer, M.U.G.; Li, B.; Cai, J.; Xu, B.; Yang, Q.; et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science* **2020**, *368*, 638–642. [CrossRef]

11. Jia, J.S.; Lu, X.; Yuan, Y.; Xu, G.; Jia, J.; Christakis, N.A. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* **2020**, *582*, 389–394. [CrossRef] [PubMed]
12. Cummings, M.J.; Baldwin, M.R.; Abrams, D.; Jacobson, S.D.; Meyer, B.J.; Balough, E.M.; Aaron, J.G.; Claassen, J.; Rabbani, L.E.; Hastie, J.; et al. Epidemiology, clinical course, and outcomes of severe and critical adults with COVID-19 in New York City: A prospective cohort study. *Lancet* **2020**, *395*, 1763–1770. [CrossRef]
13. Álvarez-Mon, M.; Ortega, M.A.; Gasulla, Ó.; Fortuny-Profitós, J.; Mazaira-Font, F.A.; Saurina, P.; Monserrat, J.; Plana, M.N.; Troncoso, D.; Moreno, J.S.; et al. A Predictive Model and Risk Factors for Case Fatality of COVID-19. *J. Pers. Med.* **2021**, *11*, 36. [CrossRef] [PubMed]
14. Moghadas, S.M.; Shoukat, A.; Fitzpatrick, M.C.; Wells, C.R.; Sah, P.; Pandey, A.; Sachs, J.D.; Wang, Z.; Meyers, L.A.; Singer, B.H.; et al. Projecting hospital utilization during the COVID-19 outbreaks in the United States. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 9122–9126. [CrossRef]
15. Mohsin, F.; Nahrin, R.; Tonmon, T.T.; Nesa, M.; Tithy, S.A.; Saha, S.; Mannan, M.; Shahjalal, M.; Faruque, M.O.; Hawlader, M.D.H. Lifestyle and Comorbidity-Related Risk Factors of Severe and Critical COVID-19 Infection: A Comparative Study Among Survived COVID-19 Patients in Bangladesh. *Infect. Drug Resist.* **2021**, *14*, 4057–4066. [CrossRef]
16. Perveen, S.; Shahbaz, M.; Keshavjee, K.; Guergachi, A. Prognostic Modeling and Prevention of Diabetes Using Machine Learning Technique. *Sci. Rep.* **2019**, *9*, 13805. [CrossRef]
17. Barra, M.; Dahl, F.A.; Vetvik, K.G.; MacGregor, E.A. A Markov chain method for counting and modelling migraine attacks. *Sci. Rep.* **2020**, *10*, 3631. [CrossRef]
18. Shan, S.; Zhao, F.; Wei, Y.; Liu, M. Disaster management 2.0: A real-time disaster damage assessment model based on mobile social media data-A case study of Weibo (Chinese Twitter). *Saf. Sci.* **2019**, *115*, 393–413. [CrossRef]
19. Deng, S.; Huang, Z.J.; Sinha, A.P.; Zhao, H. The Interaction between Microblog Sentiment and Stock Return: An Empirical Examination. *MIS Q.* **2018**, *42*, 895–918. [CrossRef]
20. Roy, A.; Nikolitch, K.; McGinn, R.; Jinah, S.; Klement, W.; Kaminsky, Z.A. A machine learning approach predicts future risk to suicidal ideation from social media data. *Npj Digit. Med.* **2020**, *3*, 78. [CrossRef]
21. Gan, C.C.R.; Feng, S.; Feng, H.; Fu, K.-W.; E Davies, S.; A Grépin, K.; Morgan, R.; Smith, J.; Wenham, C. #WuhanDiary and #WuhanLockdown: Gendered posting patterns and behaviours on Weibo during the COVID-19 pandemic. *BMJ Glob. Health* **2022**, *7*, e008149. [CrossRef] [PubMed]
22. Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **2001**, *6*, 199–231. [CrossRef]
23. Zhang, G.P. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **2003**, *50*, 159–175. [CrossRef]
24. Yu, P.; Yan, X. Stock price prediction based on deep neural networks. *Neural Comput. Appl.* **2020**, *32*, 1609–1628. [CrossRef]
25. Singh, G.; Pal, M.; Yadav, Y.; Singla, T. Deep neural network-based predictive modeling of road accidents. *Neural Comput. Appl.* **2020**, *32*, 12417–12426. [CrossRef]
26. Montoya, R.; Gonzalez, C. A Hidden Markov Model to Detect On-Shelf Out-of-Stocks Using Point-of-Sale Data. *Manuf. Serv. Oper. Manag.* **2019**, *21*, 932–948. [CrossRef]